Neural Computing and Applications manuscript No.
(will be inserted by the editor)

A framework for increasing the value of predictive data-driven models by enriching problem domain characterization with novel features

Sérgio Moro · Paulo Cortez · Paulo Rita

Received: 2015 / Accepted: 2015

Abstract The need to leverage knowledge through data mining has driven enterprises in a demand for more data. However, there is a gap between the availability of data and the application of extracted knowledge for improving decision support. In fact, more data does not necessarily imply better predictive data-driven marketing models, since it is often the case that the problem domain requires a deeper characterization. Aiming at such characterization, we propose a framework drawn on three feature selection strategies, where the goal is to unveil novel features that can effectively increase the value of data by providing a richer characterization of the problem domain. Such strategies involve encompassing context (e.g., social and economic variables), evaluating past history, and disaggregate the main problem into smaller but interesting sub-problems. The framework is evaluated through an empirical analysis for a real bank telemarketing application, with the results proving the benefits of such approach, as the area under the Receiver Operating Characteristic curve increased with each stage, improving previous model in terms of predictive performance.

Keywords Feature Selection · Decision Support · Data Mining · Telemarketing · Bank Marketing

S. Moro
ALGORITMI Research Centre, University of Minho, 4800-058 Guimarães, Portugal, and Business Research Unit (BRU-UNIDE), ISCTE - University Institute of Lisbon, 1649-026 Lisboa, Portugal
E-mail: scmoro@gmail.com

P. Cortez
Dep. of Information Systems/ALGORITMI Research Centre, University of Minho, 4800-058 Guimarães, Portugal

P. Rita
Business Research Unit (BRU-UNIDE), ISCTE - University Institute of Lisbon, 1649-026 Lisboa, Portugal

1 Introduction

In a world capable of generating exponential amounts of data, the term big data has become a common denominator for every enterprise in every industry. Such issue has been also a subject of debate in the scientific community in the past few years, generating a large research effort in both theory and practice [1]. When companies are able to overcome the initial stage of acquiring large size and high performance although somewhat still expensive data store solutions, they find themselves grappling with large amounts of data and struggling for taking real advantage of data [2, 3]. The next stage represents the current challenge for most companies, which basically is the extraction of valuable knowledge from data for effectively leveraging business decision support, whether in the form of understandable reports for explaining past events, or by translating insightful data-driven predictions into decisions that can feed operational systems. Interestingly, the later form has been foreseen by Bucklin et al. [4] in what the authors denominated as marketing decision automation systems. Wang [5] defines this stage as the move toward analytic intelligence, where advanced artificial intelligence, machine learning and data mining techniques are incorporated in decision support systems for taking the most advantage from data. In fact, data mining had a profound impact in selecting the best clients to contact, thus changing the paradigm of marketing [6, 7].

Data-driven analytical approaches are typically comprised of a knowledge discovery solution including steps such as business and data analysis, data mining on a pre-prepared dataset, and validation of results, whereas such process may be fed with recent results for providing adaptation to a changing reality, in a continuous cycle [8]. Nevertheless, successfully finding hidden patterns requires that the features characterizing each event conceal relations with the problem addressed, more specifically, with the modeled outcome features. With the advent of big data, a large number of features are usually available for modeling. Often, learning algorithms cannot by themselves straightforwardly disentangle useful data attributes (also known as features or variables) from irrelevant ones, particularly when there is a very large number of attributes that can potentially be used. Instead, algorithms will make a precious effort in finding relations between input features and the outcome to model, before finally realizing a large portion of the features are useless. Since computational time and memory requirements typically increase exponentially with the number of variables, pruning the number of input variables to a manageable number of them is mandatory. Such issue led to the emergence of an increasingly relevant branch in machine learning known as feature selection [9].

Feature selection and engineering is particularly relevant for database marketing problems [10, 11]. The creativity associated with marketing management implies that a vast number of characterizing features may influence a given problem, posing the difficult challenge of discovering them. Generic context features are known to be of high value for modeling problems through data mining. As an example, the research of Saarenpaa et al. [12] takes advantage of generic demographic indicators for distribution network planning. Moreover, associating context features that may potentially affect an instance of a problem has been shown to be highly beneficial for spatiotemporal problems [13]. The same study associates situational events for affecting the outcome of association rules. Contextual features may also be used as inputs for feeding continuously a model, as shown by our previous work [14], which pioneered the introduction of social and economic context features for improving the prediction of telemarketing contacts. Usually, real problems addressed by data mining applications encompass a temporal dimension, as the instances of the problem are occurring in different moments in time. This type of problems are typically influenced by their historic past events, as it happens in stock exchange markets [15], in retail sales [16], in fraud detection [17], and in marketing campaigns [18]. Traditionally, history information has been used for the marketing and sales domains in the form of the RFM (Recency, Frequency, Monetary) indicators [19]. However, other historic features specific of a problem can be incorporated for improving model performance, encompassing metrics for measuring customer lifetime value (LTV) [20].

It often occurs that a problem being studied is vast in its complexity, with a wide range of features influencing it in numerous ways. Data mining applications use techniques for reducing such complexity. For example, the decision tree modeling is a perfect example where this feature reduction often occurs [21]. Yet, few articles have conducted research on automated approaches for dividing the problem in smaller and more manageable sub-problems to reduce the feature selection search space [22]. Within our knowledge, there has been no such study that uses a mixed approach that combines domain expert knowledge with automated data mining techniques. The majority of research studies on feature selection focus on finding the smallest feature subset from the original set of features given a certain generalization error, while making an effort for optimizing model performance [23]. Few studies have looked at the problem of generalizing methods for extending the boundaries of the feature set beyond the original dataset that is being explored for predictive analytics purposes. Even when a large number of features are available, often happens the case that several relevant features are missing, mainly because a real world problem is affected by a myriad of variables with intrinsic relations between each other and the problem.

In this article, a framework is proposed for enriching datasets used for data mining procedures by unveiling previously unforeseen features that increase the value of the original dataset in terms of modeling performance. For validation purposes, a problem of selling long-term bank deposits through telemarketing campaigns is addressed. The results show a consistent increase in model performance, acclaiming the benefits of the suggested framework.

2 Proposed Framework and Method

The design of the experiments that ultimately led to the framework here presented focused mainly in adding value to data for data mining applications, considering data is the key ingredient for any successful data-driven project. Starting with a typical dataset for any given problem to address with a data mining approach, the emphasis is on finding previously undiscovered features that can add value to the data, and at the same time to reduce the number of features to a smaller and more manageable number for allowing computationally feasible models considering a reasonable amount of time and memory. For a proper validation of the proposed framework, some of the newly proposed features should arise in terms of feature relevance when compared to the remaining features while at the same time the model conceived must excel the performance of the baseline model without the new features. The framework designed is based on three simple but highly relevant strategies:

- Include context features;
- Evaluate past history;
- Divide and conquer strategy.

Figure 1 displays the schematic flow for finding novel features, materialized in the proposed framework. Traditional data mining projects encompass a pre-modeling stage for evaluating from the available features which are the ones that may have an impact on the problem being modeled, in a process named feature selection. The main goal of the framework presented is to extend this problem features list by adding new features for improving the prediction performance of a model based on this enriched list of features. Furthermore, some of the added features should have played a significant role in training the model. Thus another goal is to extract the relevance of some of the added features for assessing its impact.

The three strategies listed above are represented in rectangle boxes in the schematic framework. Any typical data mining problem starts with a list of features which characterize the problem, represented by series of occurrences for building the model. Then, the model is evaluated in terms of its predictive performance. The first strategy consists in assessing the context surrounding and likely affecting the problem through a domain expert. This leads to a list of newly proposed features related to problem context for enriching the initial dataset. Then, a feature reduction procedure takes place using a semi-automated method based on domain expertize and a sensitivity analysis (SA) for assessing feature relevance.

The SA measures the effect that changing the input features through their range of possible values has on the outcome of a model. Such analysis uses a pure black-box use of the data-driven models. It is based on a set sensitivity samples that are used to query the fitted models and obtain sensitivity responses. Thus, the SA approach is completely independent on the model representation and learning algorithm and it can virtually be applied to any

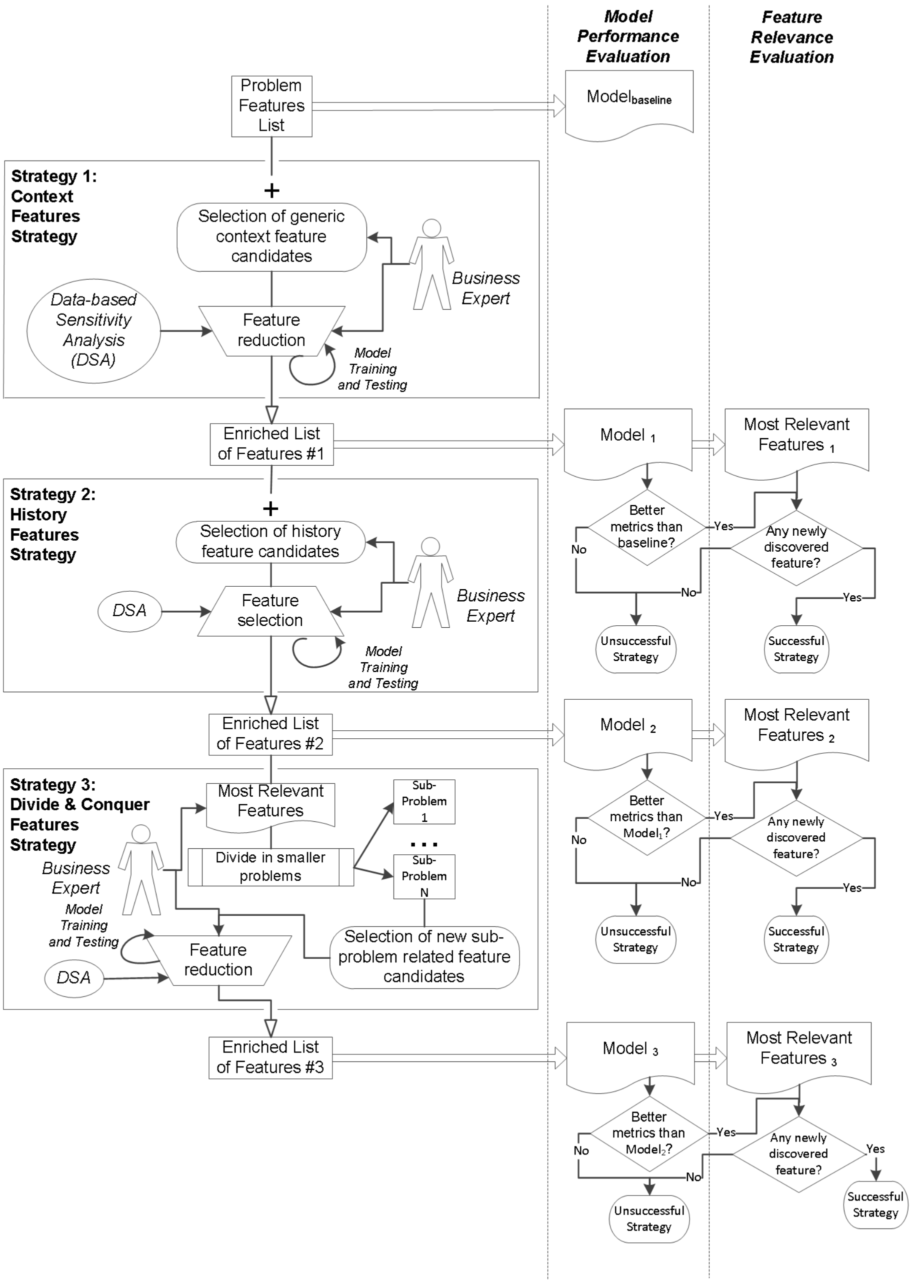

Fig. 1 Framework for finding new features

supervised learning method, including complex "black-box" models, such as neural networks and support vector machines.

The simplest SA method is termed one-dimensional SA (1D-SA). First, a baseline vector is defined with some input reference values, such as average, median or mode. Then, for each feature, L input samples are created by using all baseline values except the feature values, which are varied through its range with L levels [24]. The computational effort is given by $\mathcal{O}(M \times L \times P)$, where M denotes the total number of features and P the effort of the predict function used to build one data-driven model response. 1D-SA can only measure individual input effects and not their interactions, thus, other SA variants were proposed. The global SA (GSA) is the most complete method and uses a set of M input features that simultaneously vary with L levels [25]. GSA is computationally demanding, presenting a complexity of $\mathcal{O}(L^M \times P)$. More recently proposed [26], the data-based SA (DSA) uses N_s samples randomly selected from the training set instead of the one input baseline vector adopted by 1D-SA. Also in [26], several metrics, such as variance, gradient and a newly proposed average absolute deviation were tested for computing input relevance associated with from a set of sensitivity inputs and their respective model responses. The computational effort of DSA is much reasonable when compared with GSA: $\mathcal{O}(M \times L \times N_s \times P)$, since typically $M \times L \times N_s \ll L^M$. Moreover, DSA is capable of measuring input interactions with a performance that is similar to GSA. Taking into account these advantages, DSA was adopted in this work for all experiments. In particular, we adopted the recommendations of [26], which included using the average absolute deviation as the metric for computing input relevance. It should be further noted that the feature reduction procedure thrusts on both expert knowledge and DSA to discard irrelevant features. In this work, such approach has led to an increase in the models performance.

The enriched list of features resulted from the strategy is then used for building a model with two goals: first, assess predictive performance through the same metrics used for the model with the initial features; then, to obtain a list of the relevance of each of the features used for modeling. The strategy is considered successful in enriching the dataset if the new model outperforms the initial model and if one or more of the newly proposed features are highly ranked in terms of their input relevance. The second strategy starts with the previously enriched dataset with context features. Then, the focuses turn on analyzing the actual dataset using domain knowledge to evaluate any history information that may be extracted based on related occurrences. Using feature engineering, novel features can be designed based on the existent features of past occurrences. These features are then evaluated through a forward selection process that adds groups of related features to the previous list, assessing its impact through domain expert knowledge and the data based sensitivity analysis. The new enriched list of features is then evaluated through modeling, for measuring its impact in predictive performance and the relevance of the newly proposed features, in a procedure similar to the one of the previous strategy. Finally, the third strategy proposes to evaluate among the most relevant features in terms of DSA, which is the best candidate for splitting the problem in smaller sub-problems. Then, a domain expert decides on the feature to use for proceeding with the split. The domain expertize is again in demand for re-evaluating the sub-problem in terms of the features which better characterize this new problem. A feature reduction procedure using domain knowledge and DSA, similar to the one used in the first strategy helps to reduce the list of features to a manageable number. An evaluation approach similar to the previous strategies measures the degree of success of this strategy. The framework modular design (Figure 1) helps to contain each specific strategy in independent modules, meaning that each of them may independently be tested on any problem domain for assessing the impact on such modeling, also guaranteeing that the framework is extensible to other strategies to add features for enriching the initial dataset. Nevertheless, each of the proposed strategies uses similar techniques such as the irreplaceable domain expert knowledge, complemented with the automated DSA feature relevance evaluation.

3 Problem Description

The problem of selecting the best customers for targeting with promotional offers from an initial database is a complex task. Nevertheless, such problem is typical in the marketing domain. For the experiments presented, a bank telemarketing dataset is adopted, consisting of several telemarketing campaigns executed from May 2008 to June 2013, in a total of 52,944 phone contacts characterized by 119 features. These features encompass two different sources: (1) campaign reports, including features such as call direction (outbound or inbound) and agent gender; and (2) customer related features such as the existence of account movement blocks and customer profile assessment regarding assets, liabilities and risks. All of them were available for building the initial dataset to conduct modeling. Each campaign offered a long-term deposit from a Portuguese bank, with the result being a successful contact if the customer subscribed the deposit. While a campaign may communicate the same or a different deposit than previous campaigns, the characteristics of the deposit type are incorporated in terms of features (e.g., interest rate offered, term period) into the dataset. All calls are executed through human agents (i.e., no automated calls are made), with a large fraction of the contacts being executed in outbound, where the institution takes the initiative of contacting the client, while the remaining are inbound calls where the client is calling for any other reason and the agent gets an alert indicating the client is targeted for the campaign, giving the choice to the agent of deciding or not to approach the client for selling the deposit. A typical characteristic of targeting problems is the low successful rate. Accordingly, for the case in analysis, the dataset is unbalanced, as only 6,557 (12.38%) records are related with successes. The interest for this case has been confirmed by the research community, considering part of this dataset was published in the University of California Machine Learning Repository (http://archive.ics.uci.edu/ml/), with a high number of page hits, above a hundred thousand.

4 Results and Discussion

The bank telemarketing case study detailed in the previous section was chosen for testing the results of applying the framework. As stated in previous section, each of the three strategies is based on independent studies. As a result, the empirical analyses for the first two strategies were described in the articles of Moro et al. [14] and Moro et al. [27], while the third study relates to unpublished research. Thus, the emphasis of this paper is totally on the benefits of the application of the proposed framework. Figure 2 shows the results in three vertical layers, in the same disposition as the framework was presented in Figure 1. On the left, the results are displayed in terms of proposed features. First, a new set of suggested features related to that strategy are proposed, adding to the remaining features (the ones provided in the initial dataset for the first strategy or the features from the previously enriched dataset for the following strategies). Next, a tuning procedure adapted for each strategy leads to a reduced number of highly relevant features according to DSA, from which some of the previously proposed are among the chosen ones (the number of those is identified in the "Nr. new features included" box). In the middle layer, the metrics achieved are displayed for evaluating performance capabilities. Finally, on the right layer, the most relevant features in the top ten most relevant for the model built among those proposed in each strategy are shown, as well as the position in the top ten ranking and the relevance in percentage.

While each experiment is exhaustively explained in the individual papers, some details are highlighted. During the experiments for the first strategy for building the models, four techniques implemented in the **rminer** package [28] were tested:

- logistic regression;
- decision tree;
- support vector machine;
- ensemble of neural network.

Such comparison provided solid ground for choosing the technique that performs best. For the case of the dataset described in Section 3, the neural network clearly outperformed the remaining three techniques, as shown by Moro et al. [14]. Given these results, the ensemble of neural networks was the technique chosen for the remaining strategies.

The configuration adopted for the neural network was a multilayer perceptron with one hidden layer of H hidden nodes and one output node. A high number of hidden nodes allows learning complex nonlinear relations. The H hyperparameter sets the model learning complexity. A neural network with a value of $H = 0$ is equivalent to the LR model, while a high H value allows the neural network to learn complex nonlinear relationships. The i-th neuron (s_i) for an input $\mathbf{x}_k$ is calculated by:

$$s_i = f(w_{i,0} + \sum_{j \in P_i} w_{i,j} \times s_j) \tag{1}$$

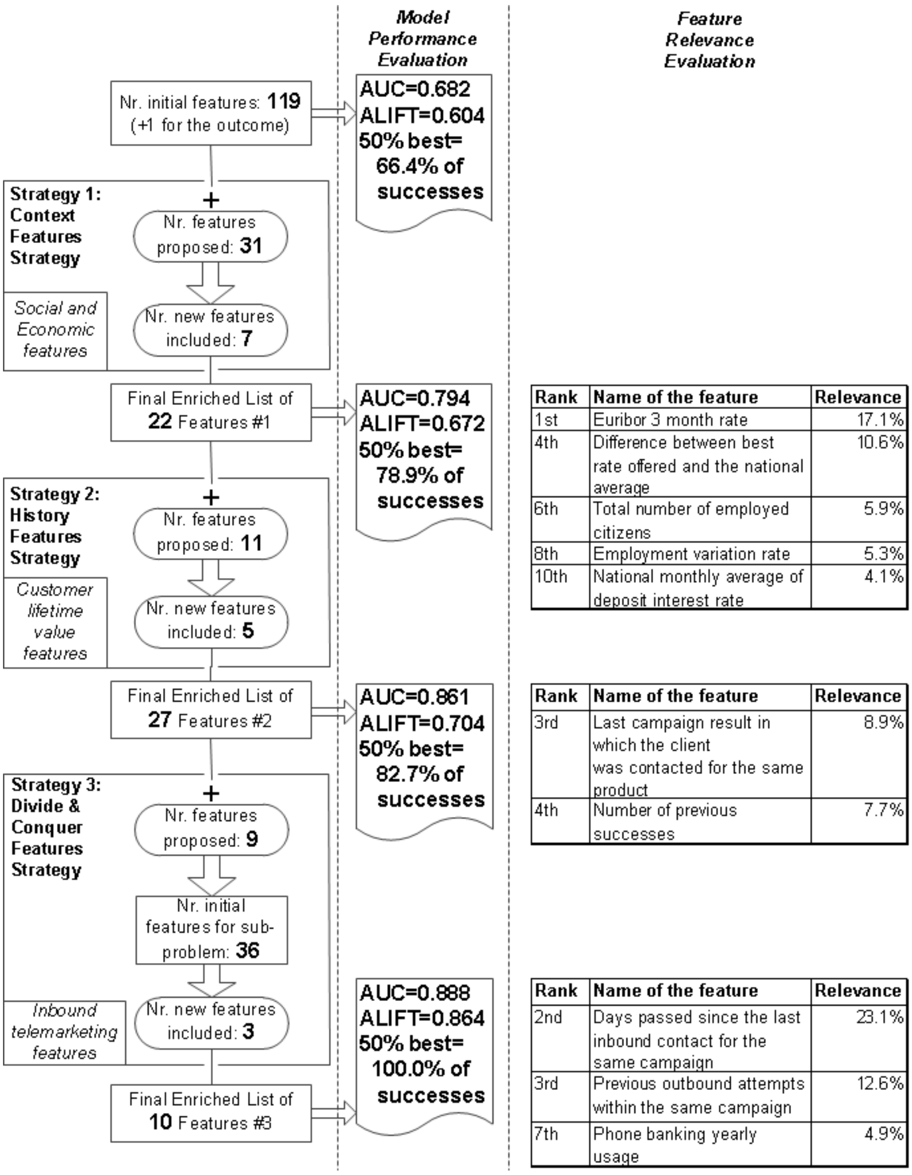

Fig. 2 Framework applied to the bank telemarketing case study

where P_i represents the set of nodes reaching node i; f is the logistic function; $w_{i,j}$ denotes the weight of the connection between nodes j and i; and $s_1 = x_{k,1}, \ldots, s_M = x_{k,M}$. The final solution is dependent of the choice of starting weights. **Rminer** addresses this problem by combining several trained networks in an ensemble [29].

For validating the models, a fixed rolling windows procedure was used to predict the outcome of the more recent year of data, from July 2012 to June 2013, including 1,293 contacts, to allow simulating a real execution of an adaptive system that keeps updating itself with the latest outcomes. Three metrics were used for measuring performance, as displayed in Figure 2:

- Area under the receiver operating characteristic curve (AUC);
- Area under the Lift cumulative curve (ALIFT);
- Selection of the 50% best clients, that is, the most likely subscribers of the deposit, using the Lift cumulative curve middle point for the selection.

The receiver operating characteristic curve shows the performance of a two class classifier across the range of possible threshold values, plotting one minus the specificity versus the sensitivity [30], while the Lift cumulative curve is a popular measure of performance in marketing applications, providing an order by dividing the dataset in ten fractions and including the most likely subscribers in the top deciles [31]. Both metrics range from 0 to 1, with the ideal model presenting metrics closer to 1. For measuring feature relevance, the DSA was used for assessing feature impact on the model. For the experiments in both the first and second strategies, the size of the sample was set to 20,000 contacts, the same number used for training the model, and the default for the **rminer** implementation of DSA (thus, $N_s = 20,000$, which assures a very robust estimation of input relevance). Nevertheless, other metrics for measuring both performance and feature relevance could be used. For the context features strategy, the hypothesis is that social and economic generic characteristics have an effect on the clients' decision to subscribe the deposits being offered through telemarketing campaigns. Thus, 31 social and economic features were proposed to the initial 119 features, resulting in a total of 150 features. The feature reduction procedure adopted included two stages: first, human domain knowledge helped in defining fourteen hypotheses materialized in fourteen groups of few features per group (e.g., the "is gender relevant?" hypothesis was characterized by three features, related with the gender of the banking agent, client and client-agent difference, 0 if same sex; 1 else); then, an automated feature reduction approach based on an adapted forward selection method based on testing the impact on the model resulted in the selection of the 22 features that maximized subscription performance, from which seven of them were chosen among the newly proposed [32]. The model built on this tuned feature set resulted in a high level of performance when compared to the initial model, materialized in an increase of the AUC from 0.682 to 0.794, an increase of the ALIFT from 0.604 to 0.672, and which can reach 78.9% of successes by selecting the half most likely subscribers from the model, whereas the initial model could only reach 66.4%. The influence of the new features on the model is emphasized by verifying that five of them were among the top ten most relevant for the model as shown on Figure 2. Most notably, the generic euribor three month rate was considered the most relevant feature, with a relevance of 17.1% for the model. Thus, for the bank telemarketing case study, the first strategy proposed was a success. This comes as a result from the mixed approach of suggesting new features using both automated feature relevance analysis through DSA and human domain knowledge. The enriched list of 22 features is used as an input for the next strategy. The history features strategy is based on a simple premise: information depth implies that data needs to be worked on so that interesting characteristics of the problem can be extracted from a dataset. Feature engineering procedures may result in previously unforeseen relevant information in terms of modeling. For marketing cases, where often attention is centered on customer interactions, evaluating customer lifetime value provides an interesting approach for extracting useful information that can be used for improving targeting models [33, 34]. Following this insight, eleven new history features were proposed, including recency, frequency and monetary (RFM) indicators [35], and past interactions in previous campaigns, such as if the client subscribed the product in another campaign. These features were computed using only the five years of data that were previously available. At the end of the feature selection process, five of those new features were included, adding to the previous 22, leading to a pruned feature set of 27. The model built increased the AUC from 0.794 to 0.861, the ALIFT from 0.672 to 0.704 and can now reach 82.7% successes by selecting the half better customers, as opposed to 78.9% with the previous model. Two of the five newly proposed and chosen features were among the ten most relevant features: the result from the previous campaign for selling a similar product was ranked in third place, and the frequency of previous successes in fourth. The improvement has less impact than previous strategy, although such result is achieved exclusively by a more in-depth analysis of the same dataset for computing new features, without the need to include extra information. Finally, the last proposed strategy focuses in dividing the initial problem in smaller and more manageable sub-problems, in an attempt to optimize feature selection for a narrower problem. Previous most relevant five features among the 27 were evaluated by a domain expert in bank telemarketing management, which led to the decision of splitting the problem and corresponding dataset in outbound and inbound telemarketing. Since most of the contacts were outbound (around 96% of them), the hypothesis is that previous features list was more representative of the outbound business. Thus, the attention turned to inbound telemarketing, arguing that the low percentage of inbound contacts (around 4%, in a total of just 1,915 of them) could have been improperly characterized. Experiments in this strategy focused solely on this reduced set of contacts, for addressing its specific characteristics. When applying the DSA method, and in order to get a robust identification of input relevance, we used 90% of the total inbound contacts, thus setting the number of selected data samples to N_s =1,724. The initial 119 features dataset was re-explored by a domain expert for finding features that are specifically relevant for inbound contacts, and propose such features in addiction to the 27 features used in previous strategy. Such procedure unveiled nine new interesting features which were proposed, resulting in a total of 36. These features characterized the new sub-problem. Next, a feature reduction procedure similar to that of the first strategy identified just ten highly relevant features, from which three of them were among the nine initially proposed. The model implemented on those ten features achieved an AUC of 0.888, as opposed to 0.862 from previous model, and an ALIFT of 0.864 (previous model reached 0.704). Nevertheless, the most impressive improvement is on the half better classified customers, which include the total number of successful contacts. From the three newly proposed features, two of them were amongst the three most relevant for the model: the number of days passed since last inbound contact for the same campaign, ranked second, and the previous outbound attempts for the same campaign, in third. The results presented for the bank telemarketing case study show the usefulness of the proposed framework, with a clear impact on the prediction of customer behavior when targeted with phone contacts for subscribing long-term deposits. Much of the success of the framework was driven by a clear human domain knowledge that resulted in interesting inputs that affected modeling, enhancing prediction performance. Although the framework is drawn on the experiments of three specific strategies, one of the most interesting aspects of the framework is its extensiveness. In fact, Figure 1 underlines each of the strategies in closed boxes, meaning that any number of iterations with other strategies could be included for finding features that could potentially enrich the dataset in terms of modeling. Also the evaluation of the impact of the new features found in each strategy is quite simple, with measurements for prediction performance, and assessment for the feature relevance to check if some of the chosen new features outstand in terms of the impact on the model.

5 Conclusions

Data-driven research is conditioned to the available data. Even in this age of big data, in a real-world environment often happens the case that the available features for characterizing the addressed domain do not cover all aspects that affect the problem domain. Such issue affects particularly marketing problems, where the creativity extends the boundaries of influential features to an unforeseen level. To address this issue, a simple and effective framework is proposed for enriching a dataset used for predictive data-driven modeling. This framework is set based on three strategies for unveiling novel features: context features, history features, and divide and conquer. Such framework was tested on a bank telemarketing problem, materializing the three strategies in novel features: social and economic features, customer lifetime value features, and specifically optimized inbound telemarketing related features. The results emphasize the impact on the specific business addressed, with the prediction performance achieving an increased number of successes, and each strategy adding benefits to previous model built based on the added features. The proposed framework transports all features used for the tuned model in the previous strategy to the next, and then tries to propose novel features according to the current strategy. In future work, we will adapt the framework to address possible interactions between the newly added features and those discarded in previous stages. For instance, the proposed framework could be compared with another that first adds enriched features and only after the third strategy it discards the irrelevant input variables. Moreover, the proposed framework could be applied to on other types of marketing problem domains to test if the results are consisted with the achieved for the bank telemarketing case study. An interesting case for applying the proposed framework would be for profiling online travel agency clients. The three strategies could result in a more tuned profiling system, considering that social and economic context highly affect the tourism and hospitality industry, for which a better evaluation of customer behavior could also provide interesting insights materialized in novel features. Furthermore, online agencies are already conducting tuned segmentation modeling of customers, with the segments serving as candidates for splitting the problem in more specific and manageable problems.

Acknowledgments

The work of P. Cortez was supported by FCT within the Project Scope UID/CEC/00319/2013. The authors would like to thank the anonymous reviewers for their helpful comments.

References

1. Nick Lee and Gordon Greenley. The primacy of data? *European journal of marketing*, 42(11/12):1141–1144, 2008.
2. Hsinchun Chen, Roger HL Chiang, and Veda C Storey. Business intelligence and analytics: From big data to big impact. *MIS quarterly*, 36(4): 1165–1188, 2012.
3. Stan Maklan, Joe Peppard, and Philipp Klaus. Show me the money: improving our understanding of how organizations generate return from technology-led marketing change. *European Journal of Marketing*, 49 (3/4):561–595, 2015.
4. Randolph Bucklin, Donald Lehmann, and John Little. From decision support to decision automation: A 2020 vision. *Marketing Letters*, 9(3):235–246, 1998.
5. Fei-Yue Wang. A big-data perspective on ai: Newton, merton, and analytics intelligence. *Intelligent Systems, IEEE*, 27(5):2–4, 2012.
6. Morris B Holbrook and James M Hulbert. Elegy on the death of marketing: Never send to know why we have come to bury marketing but ask what you can do for your country churchyard. *European Journal of Marketing*, 36(5/6):706–732, 2002.
7. Babak Sohrabi, Payam Mahmoudian, and Iman Raeesi. A framework for improving e-commerce websites usability using a hybrid genetic algorithm and neural network system. *Neural Computing and Applications*, 21(5): 1017–1029, 2012.

8. Zbigniew Michalewicz, Martin Schmidt, Matthew Michalewicz, and Constantin Chiriac. Case study: an intelligent decision support system. *Intelligent Systems, IEEE*, 20(4):44–49, 2005.
9. Hongying Liu, Edward Dougherty, Jennifer G Dy, Kari Torkkola, Eugene Tuv, Hua Peng, Chibiao Ding, Fuhui Long, Michael Berens, Lowell Parsons, et al. Evolving feature selection. *Intelligent systems, IEEE*, 20(6): 64–76, 2005.
10. Ronen Meiri and Jacob Zahavi. Using simulated annealing to optimize the feature selection problem in marketing applications. *European Journal of Operational Research*, 171(3):842–858, 2006.
11. Mário Teixeira Reis Neto, Jeislan Carlos de Souza, and Gustavo Queiroga Souki. Identifying variables that predict clients' propensity to end their checking accounts. *International Journal of Bank Marketing*, 29 (6):489–507, 2011.
12. Jukka Saarenpaa, Mikko Kolehmainen, Matti Mononen, and Harri Niska. A data mining approach for producing small area statistics-based load profiles for distribution network planning. In *Industrial Technology (ICIT), 2015 IEEE International Conference on*, pages 1236–1240. IEEE, 2015.
13. Muhammad Shaheen, Muhammad Shahbaz, and Aziz Guergachi. Context based positive and negative spatio-temporal association rule mining. *Knowledge-Based Systems*, 37:261–273, 2013.
14. Sérgio Moro, Paulo Cortez, and Paulo Rita. A data-driven approach to predict the success of bank telemarketing. *Decision Support Systems*, 62: 22–31, 2014.
15. Guo-Dong Li, Daisuke Yamaguchi, and Masatake Nagai. The development of stock exchange simulation prediction modeling by a hybrid grey dynamic model. *The International Journal of Advanced Manufacturing Technology*, 36(1-2):195–204, 2008.
16. Chern Ching-Chin, Ao Ieong Ka Ieng, Wu Ling-Ling, and Kung Ling-Chieh. Designing a decision-support system for new product sales forecasting. *Expert Systems with applications*, 37(2):1654–1665, 2010.
17. Koosha Golmohammadi and Osmar R Zaiane. Data mining applications for fraud detection in securities market. In *Intelligence and Security Informatics Conference (EISIC), 2012 European*, pages 107–114. IEEE, 2012.
18. Vera L Miguéis, Ana S Camanho, and João Falcão e Cunha. Customer data mining for lifestyle segmentation. *Expert Systems with Applications*, 39(10):9359–9366, 2012.
19. Ching-Hsue Cheng and You-Shyang Chen. Classifying the segmentation of customer value via rfm model and rs theory. *Expert systems with applications*, 36(3):4176–4184, 2009.
20. I-Cheng Yeh, King-Jang Yang, and Tao-Ming Ting. Knowledge discovery on rfm model using bernoulli sequence. *Expert Systems with Applications*, 36(3):5866–5871, 2009.
21. J. Ross Quinlan. Learning decision tree classifiers. *ACM Computing Surveys (CSUR)*, 28(1):71–72, 1996.

22. Y Liu and M Schumann. Data mining feature selection for credit scoring models. *Journal of the Operational Research Society*, 56(9):1099–1108, 2005.
23. Jorge R Vergara and Pablo A Estévez. A review of feature selection methods based on mutual information. *Neural Computing and Applications*, 24 (1):175–186, 2014.
24. Robert H Kewley, Mark J Embrechts, and Curt Breneman. Data strip mining for the virtual design of pharmaceuticals with neural networks. *Neural Networks, IEEE Transactions on*, 11(3):668–679, 2000.
25. Paulo Cortez and Mark J Embrechts. Opening black box data mining models using sensitivity analysis. In *Computational Intelligence and Data Mining (CIDM), 2011 IEEE Symposium on*, pages 341–348. IEEE, 2011.
26. Paulo Cortez and Mark J Embrechts. Using sensitivity analysis and visualization techniques to open black box data mining models. *Information Sciences*, 225:1–17, 2013.
27. Sérgio Moro, Paulo Cortez, and Paulo Rita. Using customer lifetime value and neural networks to improve the prediction of bank deposit subscription in telemarketing campaigns. *Neural Computing and Applications*, 26(1): 131–139, 2015.
28. Paulo Cortez. Data mining with neural networks and support vector machines using the r/rminer tool. In *Advances in data mining. Applications and theoretical aspects*, pages 572–583. Springer, 2010.
29. T. Hastie, R. Tibshirani, and J. Friedman. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction.* Springer-Verlag, NY, USA, 2nd edition, 2008.
30. Tom Fawcett. An introduction to roc analysis. *Pattern recognition letters*, 27(8):861–874, 2006.
31. David S Coppock. Why lift? data modeling and mining. *Information Management*, pages 5329–1, 2002. URL `http://www.information-management.com/news/5329-1.html`. [Online; accessed 19-July-2013].
32. Isabelle Guyon and André Elisseeff. An introduction to variable and feature selection. *The Journal of Machine Learning Research*, 3:1157–1182, 2003.
33. Leilei Tang, LC Thomas, S Thomas, and JF Bozzetto. It's the economy stupid: modelling financial product purchases. *International Journal of Bank Marketing*, 25(1):22–38, 2007.
34. Chia-Chen Chen. Rfid-based intelligent shopping environment: a comprehensive evaluation framework with neural computing approach. *Neural Computing and Applications*, 25(7-8):1685–1697, 2014.
35. Kristof Coussement, Filip AM Van den Bossche, and Koen W De Bock. Data accuracy's impact on segmentation performance: Benchmarking rfm analysis, logistic regression, and decision trees. *Journal of Business Research*, 67(1):2751–2758, 2014.